Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models

Year: 2023 | Citations: 747 | Authors: Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang

---

Abstract

ChatGPT is attracting a cross-field interest as it provides a language interface with remarkable conversational competency and reasoning capabilities across many domains. However, since ChatGPT is trained with languages, it is currently not capable of processing or generating images from the visual world. At the same time, Visual Foundation Models, such as Visual Transformers or Stable Diffusion, although showing great visual understanding and generation capabilities, they are only experts on specific tasks with one-round fixed inputs and outputs. To this end, We build a system called \textbf{Visual ChatGPT}, incorporating different Visual Foundation Models, to enable the user to interact with ChatGPT by 1) sending and receiving not only languages but also images 2) providing complex visual questions or visual editing instructions that require the collaboration of multiple AI models with multi-steps. 3) providing feedback and asking for corrected results. We design a series of prompts to inject the visual model information into ChatGPT, considering models of multiple inputs/outputs and models that require visual feedback. Experiments show that Visual ChatGPT opens the door to investigating the visual roles of ChatGPT with the help of Visual Foundation Models. Our system is publicly available at \url{https://github.com/microsoft/visual-chatgpt}.